From Gene Expression modelling to CoRegulation networks for Arabidopsis

Rim ZAAG
PhD student
Bioinformatics for Predictive Genomics Team

Thesis supervisors
Marie-Laure Martin-Magniette
Etienne Delannoy

NETBIO
Thursday 18th September 2014

Context and Background

- A huge ‘orphan of function’ gene space:
  - 20% to 40% of the predicted genes for completely sequenced eukaryotic organisms have no assigned function (Hanson et *al*., 2010).
  - More than 5000 Arabidopsis genes are still hypothetical or unknown genes according the TAIR v10 annotation.
- Our knowledge about gene candidates involved in the adaptation of plants to their environment remains partial (of potential interest for crop improvement).

Context and Background

- Functional annotation procedures based on sequence similarities have reached their limitations

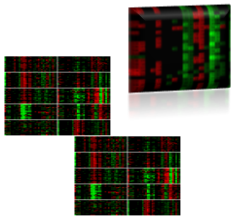
- One gene-one enzyme hypothesis is now considered as an oversimplification

- Availability of thousands of transcriptomes: allow us to shift from a 'gene by gene' approach to more global approach through the 'guilt by association' concept.

- **Hypothesis: Coexpressed genes have likely related biological functions** (Eisen et al., 1998)

Classical Flowchart

Gene Expression Omnibus



CoExpression analysis

Bioinformatic characterisation
(GO, TFBS, PPI,TF-target, bibliome…)

Focus on specific genes for biological validation

Drawbacks

- Data are generally extracted from international repositories
- It leads to heterogeneous data in terms of acquisition and preprocessing.
- Coexpression generally done by analyzing gene pairs (Pearson correlation)
- It is a local point of view of a complex question.

Our Approach: Goals & methods

Goals:

I. Provide a global overview of the coexpression units of genes responding to a panel of stress stimuli in *Arabidopsis thaliana*.
II. Go beyond the coexpression to identify coregulated modules of functional partner genes.
III. Inference of function to orphan genes in well-characterized modules.

Methods: original features

- The specificities of the dataset and the biological theme: homogeneous and dedicated transcriptomic data.
- The method of clustering: model based method
- The integration of various resources to improve the functional inference.

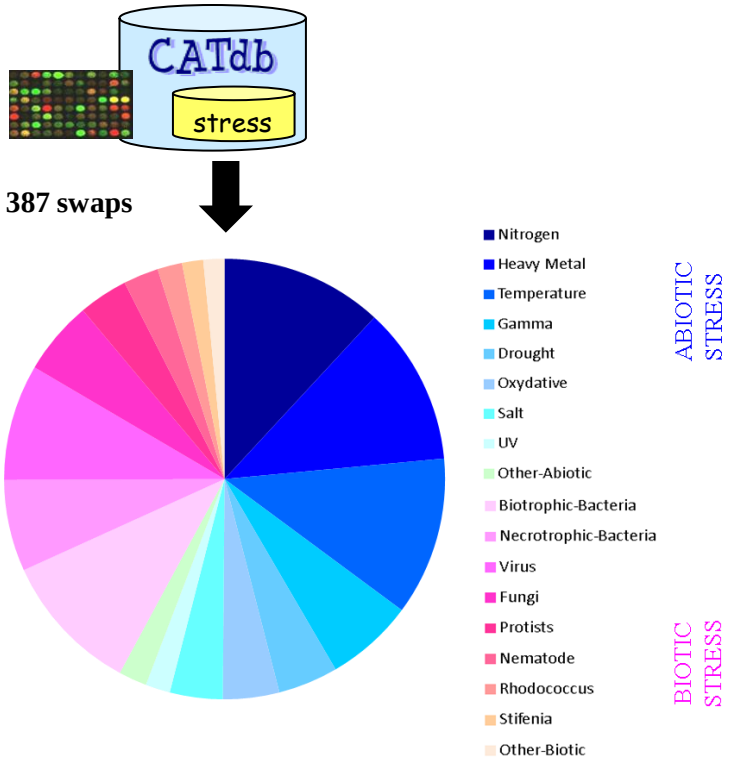
Transcriptomic Dataset

CATdb :Complete Arabidopsis transcriptome database

- Large and homogeneous transcriptome resource generated by the CATMA platform of URGV and available in CATdb (http://urgv.evry.inra.fr/CATdb; Gagnot et al., NAR 2008).

- ~ 6000 original genes not present in the commonly used ATH1 Affymetrix DNA chip.

- All experiments dedicated to stresses were considered: 9 biotic and 9 abiotic stress categories



Differential expression analysis → **17 264** genes have transcription 'impacted' (directly or not) by at least one stress experiment

CoExpression Analysis

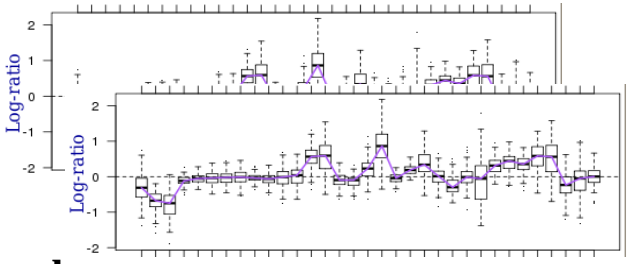
Gene Clustering: Identification of co-expressed genes from the expression differences through a Model based clustering method for each stress category.

Matrix
{ genes x experiments}
By stress

Gaussian Mixture Model

Mathematical Criterion to select the cluster number (BIC)

Classification rule based on conditional probabilities

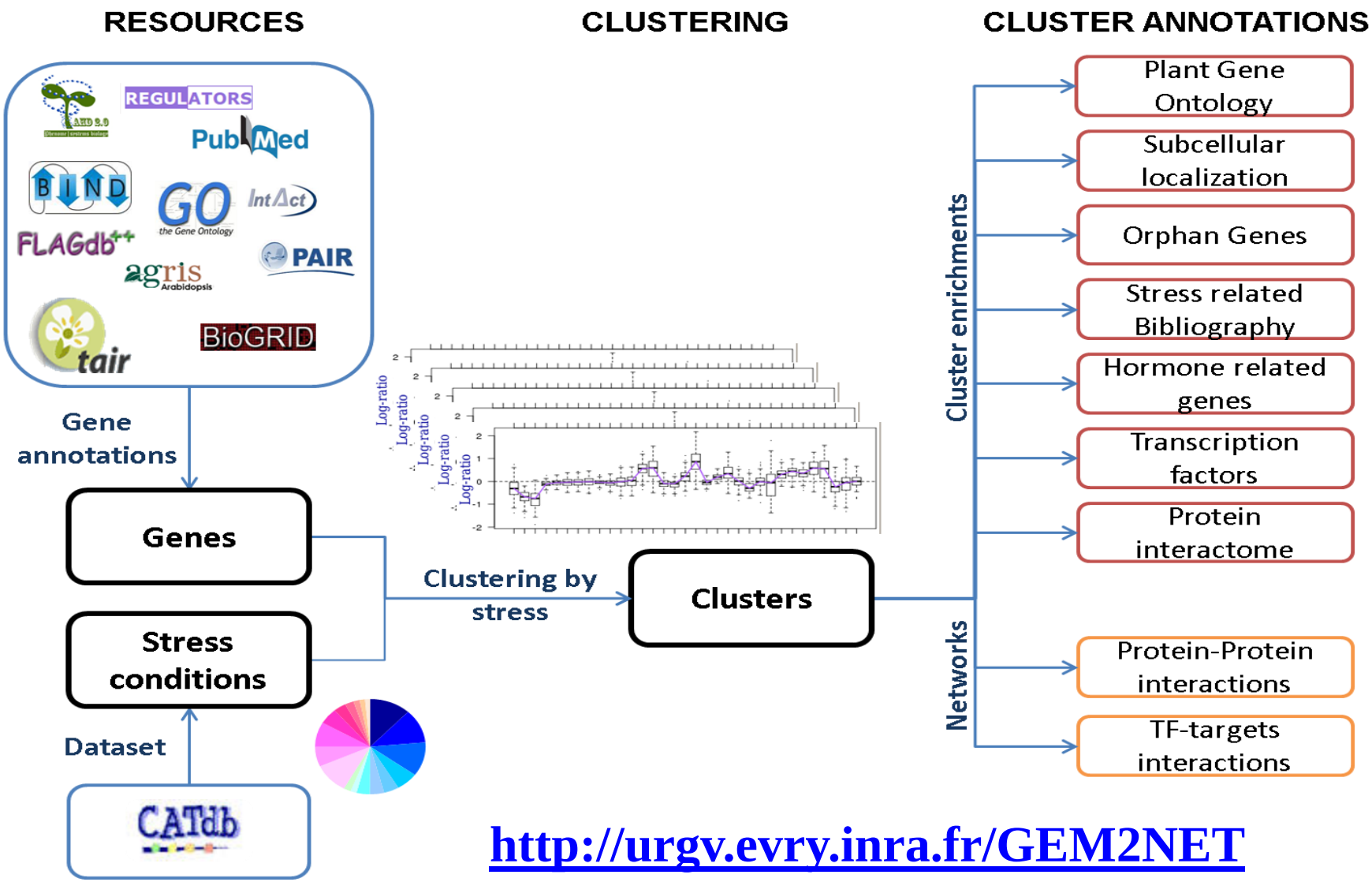


~700 Clusters of Coexpressed Genes

Stress category	Gene_nb	Clusters_nb
Nitrogen	13 495	59
Temperature	11 365	34
Drought	8 143	34
Salt	5 729	30
Heavy metal	10 617	57
UV	7 894	37
Gamma	5 350	32
Oxydative stress	10 127	52
Nectrophic bacteria	11 220	50
Biotrophic bacteria	12 023	56
Fungi	9 773	51
Rhodococcus	1 900	13
Oomycete	5 508	31
Nematode	7 413	27
Stifenia	1 525	17
Virus	11 832	54

GEM2Net Flowchart:

This project has been implemented as a new CATdb module: GEM2Net associated with a user-friendly Interface

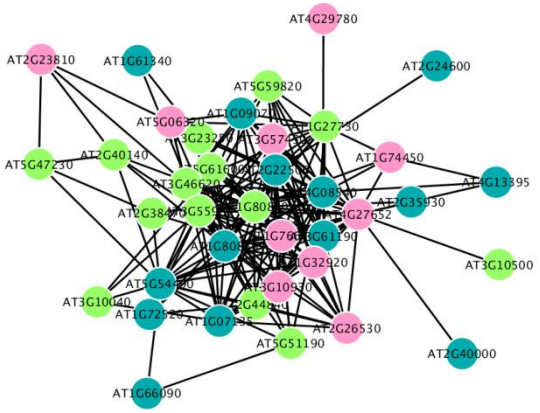


http://urgv.evry.inra.fr/GEM2NET

Functional inference by coregulation analysis

CoExpression clusters for each category of stress

Integration



CoRegulation Network

Identification of Coregulated genes → **Describe groups of functional partners** → **Annotation of orphan genes**

CoRegulation Analysis

What is CoRegulation?

Group of genes which are coexpressed in several stress conditions : key players of stress response

CoRegulation Analysis

What is CoRegulation?

Group of genes which are coexpressed in several stress conditions : key players of stress response

How to perform it?

CoRegulation Analysis

What is CoRegulation?

Group of genes which are coexpressed in several stress conditions : key players of stress response

How to perform it?

Occurrence of pairs of coexpressed genes conserved in several stresses among the 18 considered stress categories

CoRegulation Analysis

What is CoRegulation?

Group of genes which are coexpressed in several stress conditions : key players of stress response

How to perform it?

Occurrence of pairs of coexpressed genes conserved in several stresses among the 18 considered stress categories

Pairs conserved in at least n stresses

n	Nbr pair of genes
2	5 533 013
3	423 771
4	68 875
5	19 113
6	6 987
7	3 366
8	1 679
9	786
10	324
11	171
12	81
13	39
14	12
15	6

CoRegulation Analysis

What is CoRegulation?

Group of genes which are coexpressed in several stress conditions : key players of stress response

How to perform it?

Occurrence of pairs of coexpressed genes conserved in several stresses among the 18 considered stress categories

When coexpression becomes coregulation?

Pairs conserved in at least n stresses

n	Nbr pair of genes
2	5 533 013
3	423 771
4	68 875
5	19 113
6	6 987
7	3 366
8	1 679
9	786
10	324
11	171
12	81
13	39
14	12
15	6

CoRegulation Analysis

➢ We searched a threshold k at which the probability that a pair of genes is coexpressed in k stresses is significantly different from random.

➢ Permutation scheme:

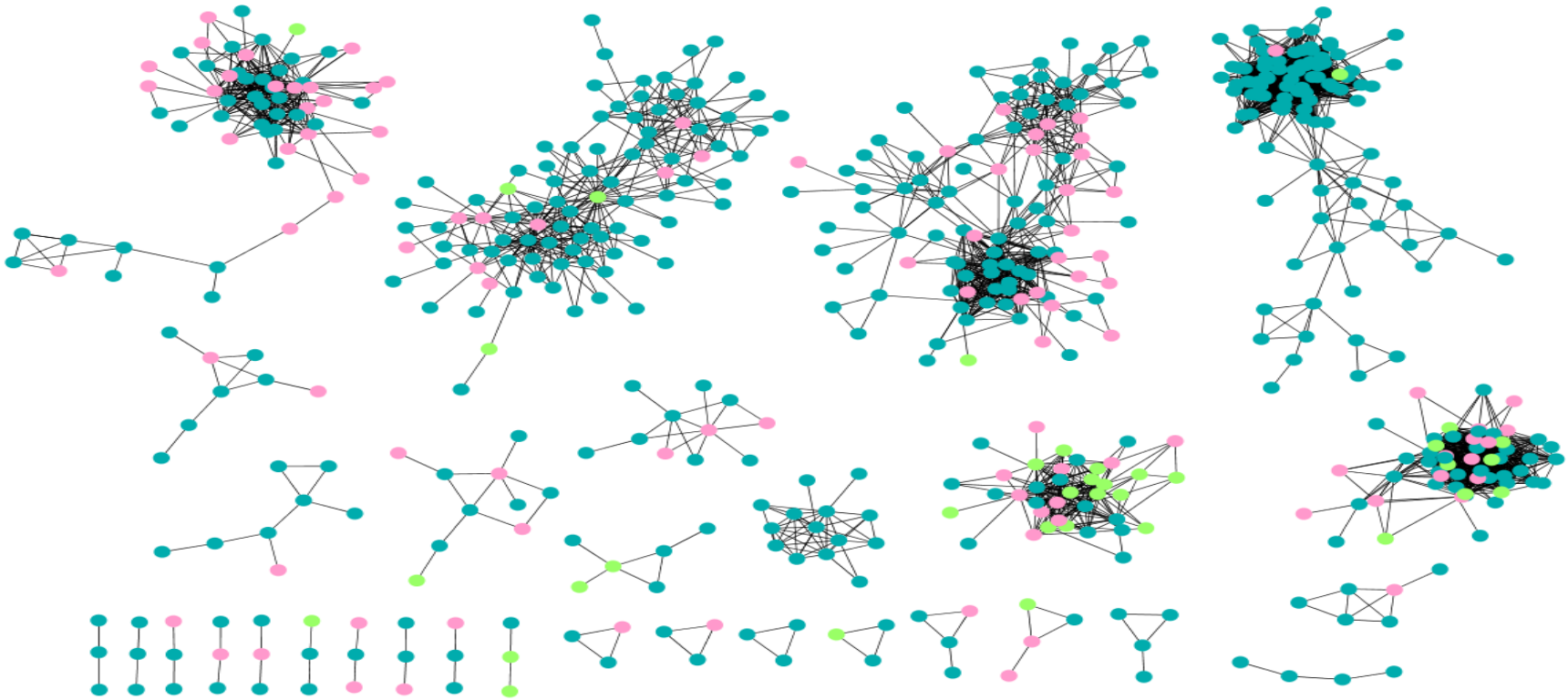
1. Do 1000 times
    a) Shuffle gene classification within each stress category.
    b) Occurrence calculation
2. Error rate calculation: average of occurrence in random samples divided by the occurrence in our data.

Number of stresses (n)	Random Network	Biological Network	Error_rate
4+	1 549	32 313	4.79%
5+	12	13 200	0.09%
6+	0	6 216	0%
7+	0	3 366	0%

Occurrence of pairs conserved in at least n stresses within our random and biological networks

CoRegulation Networks

Pairs conserved in at least 7 stresses
867 genes, 3366 pairs



Legend

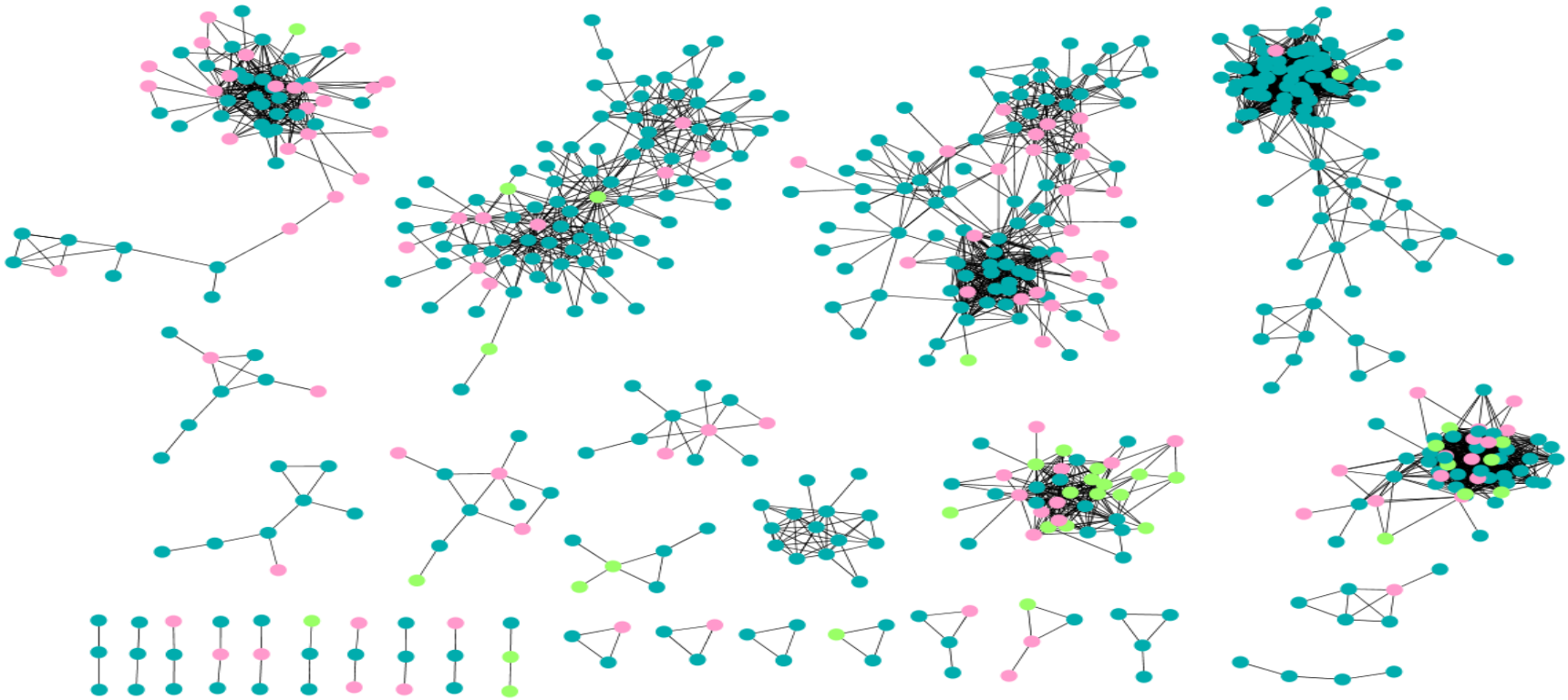
- Coregulated genes
- Orphan genes
- TF

CoRegulation Networks

Pairs conserved in at least 7 stresses
867 genes, 3366 pairs

contains

178 Orphans
57 TF



Legend

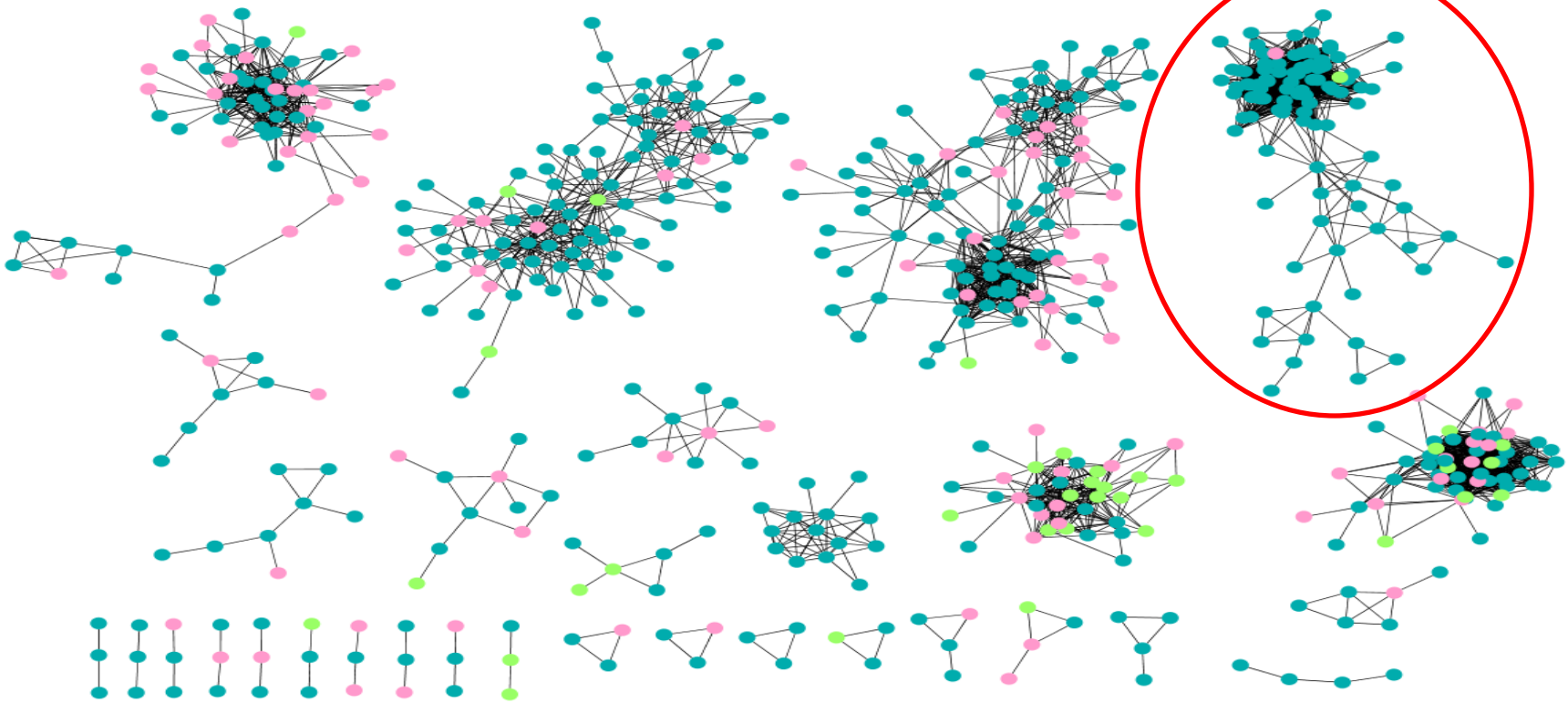
- Coregulated genes
- Orphan genes
- TF

CoRegulation Networks

Pairs conserved in at least 7 stresses
867 genes, 3366 pairs

contains

178 Orphans
57 TF



31 connected components

Legend

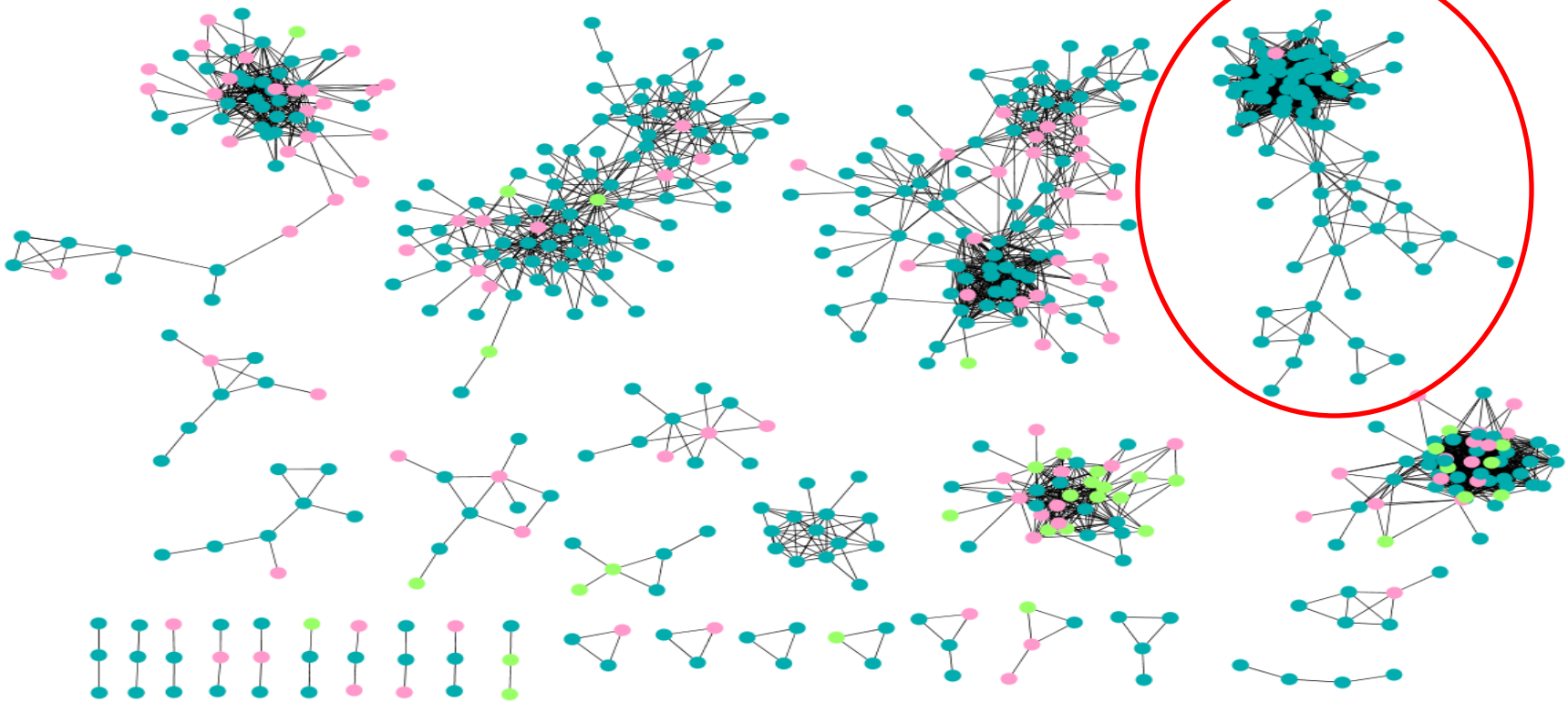
- Coregulated genes
- Orphan genes
- TF

CoRegulation Networks

Pairs conserved in at least 7 stresses
867 genes, 3366 pairs

contains

178 Orphans
57 TF

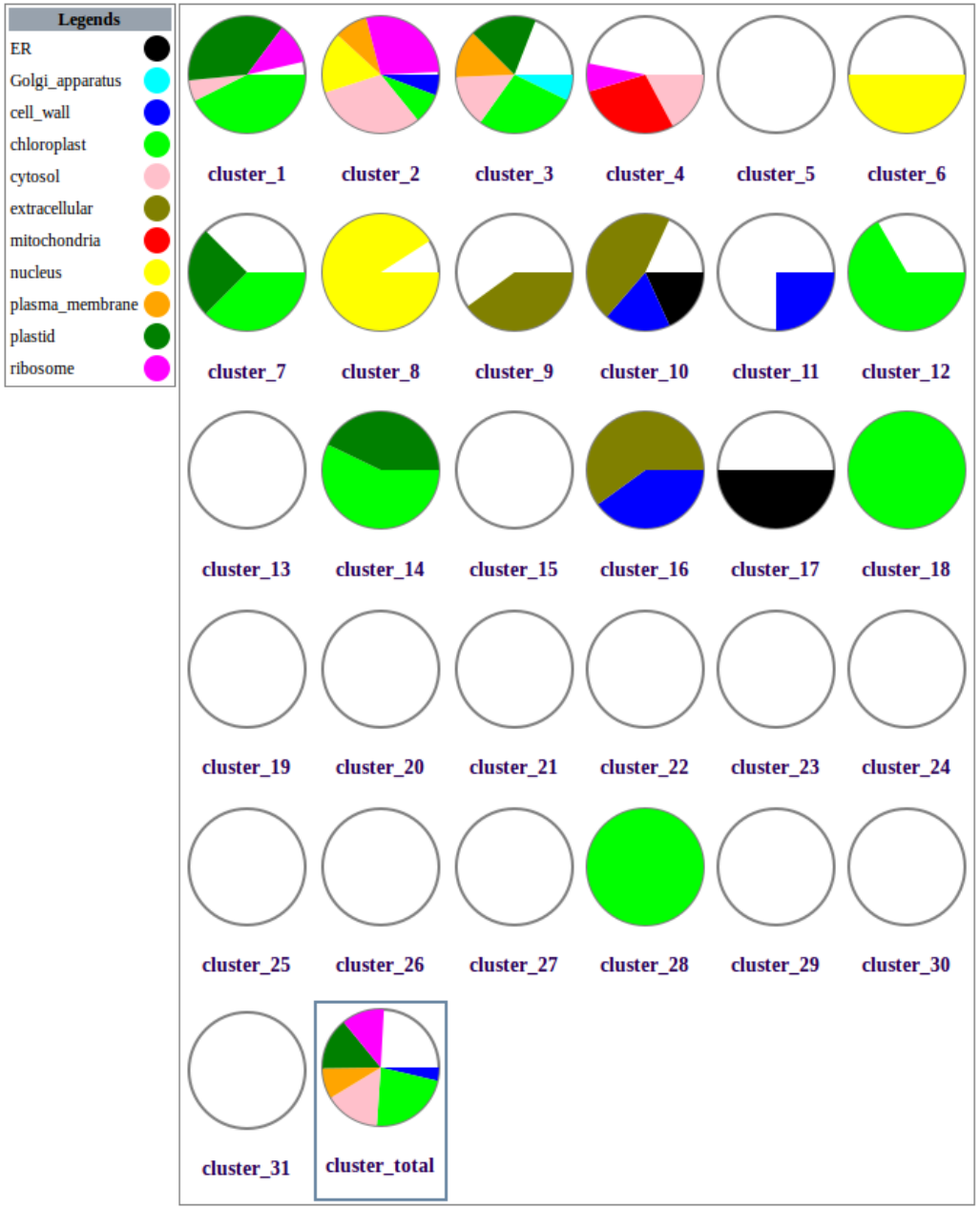


Legend

- Coregulated genes
- Orphan genes
- TF

31 connected components
Functional Modules ???

Gene Ontology Enrichment Analysis

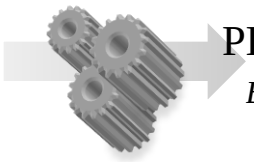


GO:Cellular Component Ontology

Specific and homogeneous modules

Cis-regulatory motifs Enrichment Analysis

Global CoRegulation Network Analysis
(867 genes)



PLMDETECT
Bernard et al., 2010

- 30 TFBS are found Over-represented by comparison with the whole genome present at most in 30% of promoters.

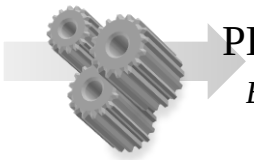
CoRegulation Network Analysis by component
(9 largest components)

PLMDETECT
Bernard et al., 2010

- 8 components are enriched in TFBS.
- 4 components are enriched with a pattern that is present in over 60% of their promoters

Cis-regulatory motifs Enrichment Analysis

Global CoRegulation Network Analysis
(867 genes)



PLMDETECT
Bernard et al., 2010

- 30 TFBS are found Over-represented by comparison with the whole genome present at most in 30% of promoters.

CoRegulation Network Analysis by component
(9 largest components)

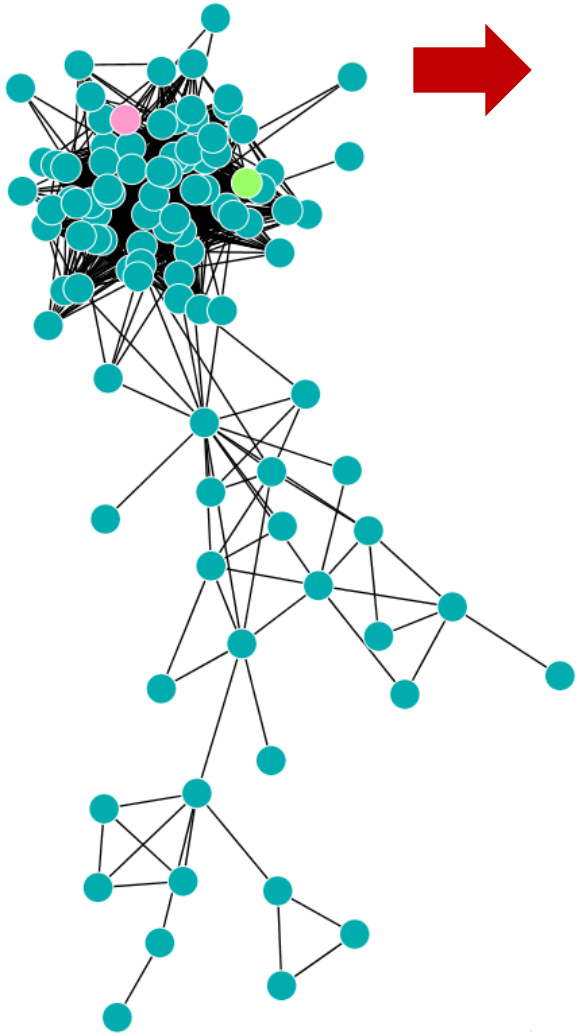
PLMDETECT
Bernard et al., 2010

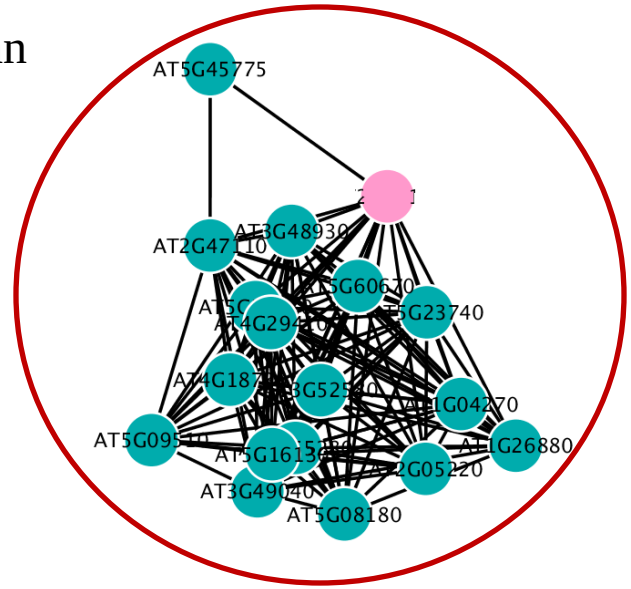
- 8 components are enriched in TFBS.
- 4 components are enriched with a pattern that is present in over 60% of their promoters

- **Genes are under the control of the same common regulators**
- **Using the network topology is a good track for identifying modules**
- **Module = Functional partners**

Example of an orphan identification within a module

98 genes, 1112 pairs, 1 orphan

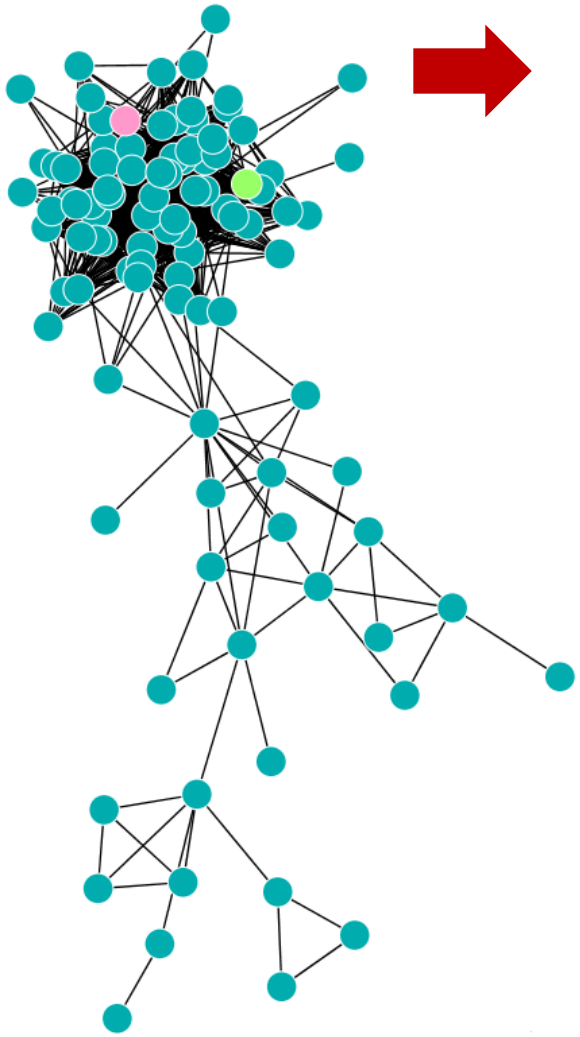


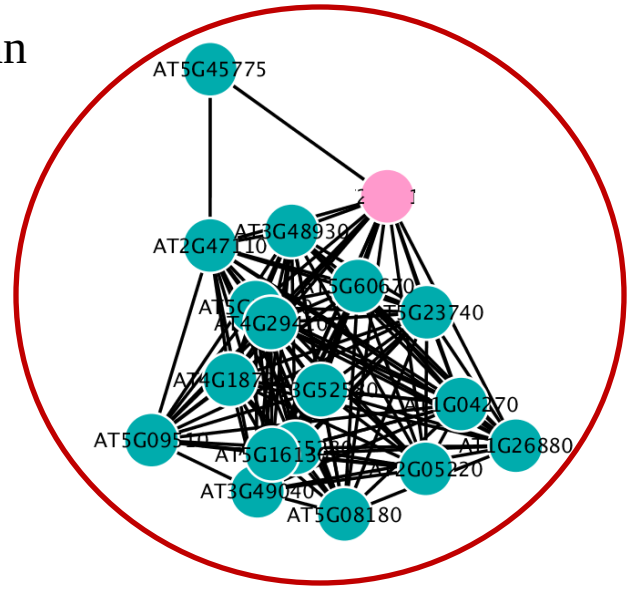


First Neighbors of the orphan gene

Example of an orphan identification within a module

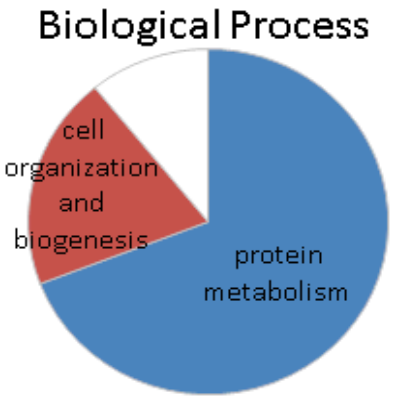
98 genes, 1112 pairs, 1 orphan

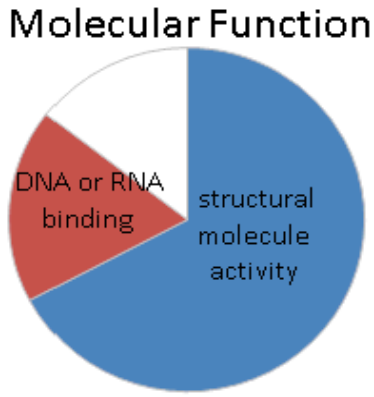




First Neighbors of the orphan gene

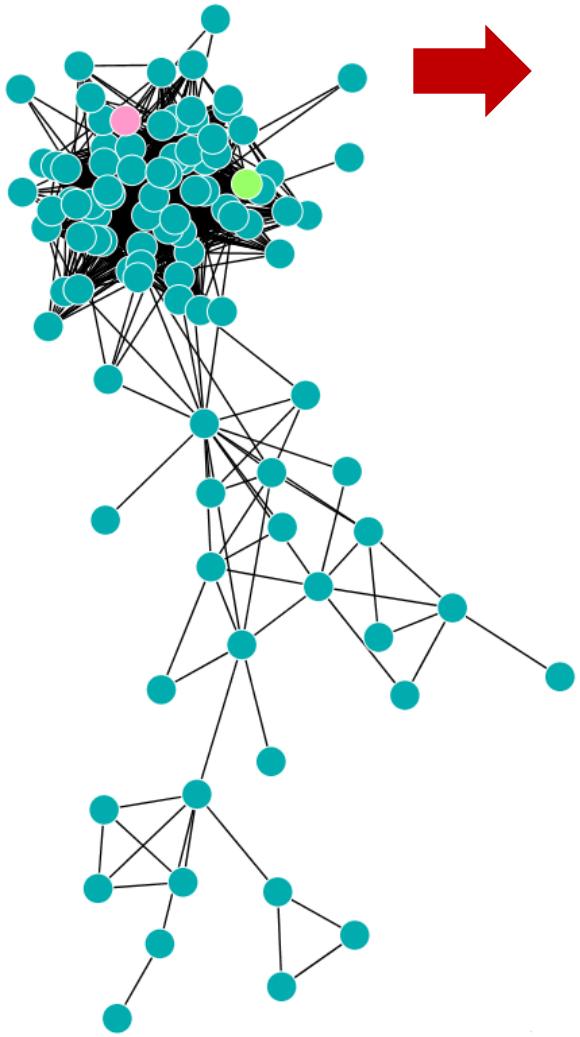
15 of the 17 neighbors are annotated as « Structural constituent of ribosome »

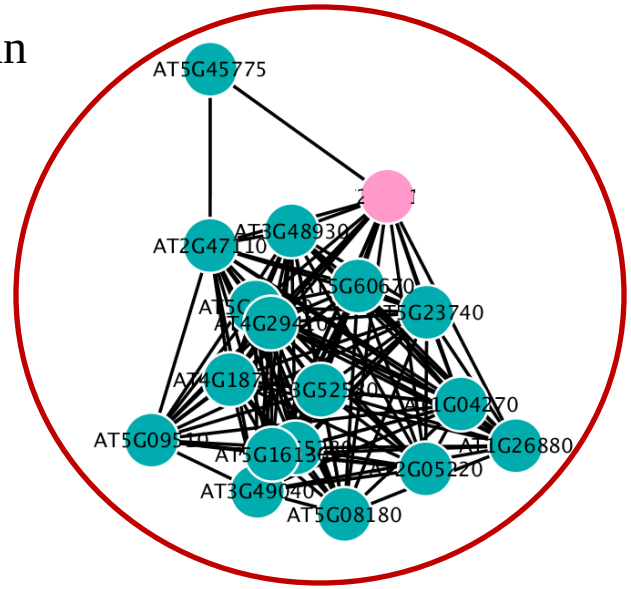




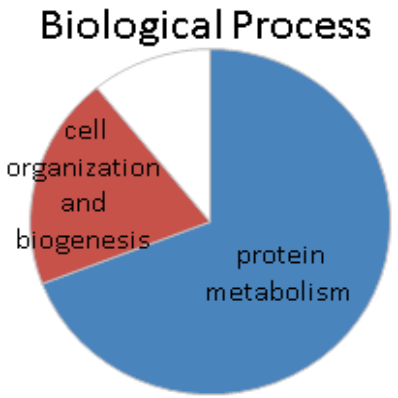
Example of an orphan identification within a module

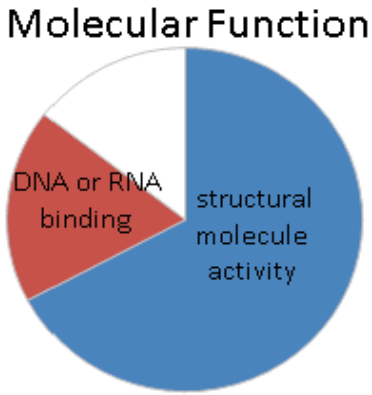
98 genes, 1112 pairs, 1 orphan





First Neighbors of the orphan gene





15 of the 17 neighbors are annotated as « Structural constituent of ribosome »

This orphan gene most likely codes for a ribosomal protein

Conclusion & Prospects

At Network « **7+** »
stresses scale

Conclusion & Prospects

At Network « **4+** »
stresses scale

Conclusion & Prospects

At Network « **4+** »
stresses scale

Identification of Coregulated genes → **Describe groups of functional partners** → **Annotation of orphan genes**

Conclusion & Prospects

At Network « **4+** »
stresses scale

Identification of Coregulated genes → **Describe groups of functional partners** → **Annotation of orphan genes**

Topolgical Network Analysis Method

✓ Refine the search of functional modules using methods of topological network analysis

✓ Integration with interactome data to improve the quality of function inference

✓ First step toward regulatory networks

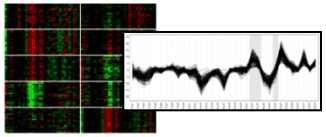
Acknowledgements

URGV

Bioinformatics

ML.Martin-Magniette
S. Aubourg
V. Brunaud
G. Rigaill
J.-P. Tamby
C.Guichard
Z.Tariq

Model-based clustering



G. Celeux (Orsay)
C. Maugis (Toulouse)

INRIA

Transcriptome, mutants and stress

E.Delannoy
C. Lurin, S.Balzergue

INRA AgroParisTech cnrs

Funding : AllEnvi INRA

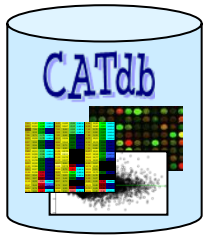
GAP
BV
MIA

Thank you for your attention

Next steps : From clusters to gene networks

From co-regulation clusters to groups of putative functional partners
- Merging and extension of clusters (functions, profiles),
- Experimental step (validation): Transcriptomes of mutants (putative regulators and TF will be targeted)

Transcriptome platform

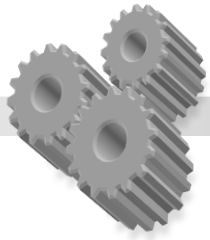


Putative functional partners

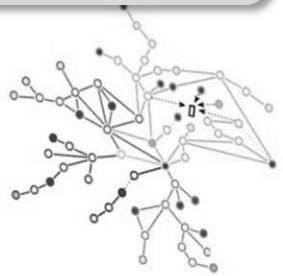
experiment selection

Gaussian Graphical Model

MIA INRA université évry val-d'essonne



Giraud et al. (2009)



Prediction of gene networks (conditional dependencies)

New candidate genes for crop improvement

Translational research

TILLing platform